# The future of science arXiving

Paul Ginsparg shares his thoughts about the future of the preprint server he created 25 years ago.

**The arXiv was born at a time when computer technologies such as e-mail and the web were taking off. Do you think that deep learning and big data analytics are now likely to take the arXiv to another stage?**
Today the combination of 25 years of longitudinal usage data with over a million articles, actively tagged and curated by submitters and moderators, constitutes a treasure trove of data for machine learning algorithms. In recent years we have experimented with recommendation systems based on both textual and usage data, with improving the article navigation based on a language model for the texts, and with automating the mechanisms for quality control. For instance, I built a system to assist and ultimately predict moderator actions. It grew increasingly accurate for category adjustments and identifying low-quality submissions, such that the overall system has now become substantially dependent on it to supplement the moderators. It also instantly analyses the full text of every submission, something we could never ask of volunteer moderators seeing 5–10 submissions per day. And it has complete data on all past moderator and submitter interactions, and various other features that are not necessarily at moderator's fingertips. As frequently happens with these 'big data' machine learning systems, they get to a point where it is difficult to understand how they work so well, but then we can stop worrying about it.

**What about social networking tools? The recent arXiv user survey revealed limited interest in such functionalities.**
For many it was more like active opposition to having such tools directly on the main site. Well over 20 years ago, we had considered comments and numerical rating systems, but received unambiguous feedback to remain focused on the basic dissemination task. Part of the issue was that contentious comments would have to be moderated, requiring human labour and reducing the scalability of the system. Ultimately it was decided that such facilities should be outsourced, and it remains a good decision.

About a decade ago, when blogging became popular, we experimented with 'trackbacks' linked from the abstract pages, providing a distributed means of moderating the discussions. Now there are social networking platforms for everything, and trolls and flame wars need to be policed. So for the time being it makes sense to continue to piggyback on existing external services. We do see in the activity logs how readers are referred to the arXiv by various social media platforms, and there is a fascinating interplay between those and the conventional news media drawing large numbers of researchers and members of the general public.

JOHN D. & CATHERINE T. MACARTHUR FOUNDATION

**What is your vision for the arXiv of the future?**
Much of the basic methodology (as well as some of the original software) has withstood the test of time, so we could imagine that 'the future will be much like the present, only longer'. On the other hand, I am on record as suggesting over 20 years ago that the current metastable state in scholarly publications, of preprint servers coexisting with conventional online publications, could not possibly persist long beyond the year 2000 — the argument is still probably correct, but the calculation of the time constant needs to be refined.

In recent years, however, there has been growth in preprint usage in other subject areas, with researchers taking more control of when and how their research results are announced.

The asapbio.org movement has recently taken a top-down approach to encouraging researchers in biology to take similar control, and a few preprint servers have been slowly growing. Owing to the structural constraints imposed by the arXiv's embedding in the Cornell University Library, it probably makes sense for other disciplines to be supported by a distributed network, perhaps endorsed (and supported) by funding agencies. To build a global knowledge network we just need a comprehensive index to long-term stable repositories, including the arXiv, with some compatible interface.

As for the arXiv itself, I would hope that there will be various forms of increased personalization, with users better able to tailor customized views and content alerts. I have, for example, a simple proposal for submitters to be able to curate an area of the abstract page to provide links to associated resources (linked data, code, video presentations, blog comment threads, slides and so on), which would improve the interoperability with a variety of services (including new overlay journals or semantic overlays to gain visibility and traction). Because it shows up as strongly desired by the users we surveyed, perhaps it will make it to the top of the priority list before too long. But this is just scratching the surface of the possibilities.

**The arXiv has huge potential for knowledge discovery. Can we imagine an arXiv Watson helping physicists?**
It is likely to happen, whether provided by the arXiv or by some third-party overlay. The language model mentioned above is a first step to automated parsing for meaning. With more sophisticated machine learning algorithms in tandem with authoring tools to produce better-structured articles, we can imagine significant improvements in data mining for meaning, providing the basis for an expert system.

A more detailed semantic representation would also feed into the above-mentioned personalization for readers, together with the extensive usage data for collaborative filtering, co-author and citation networks, among other features. Some of this I am already using for my own purposes, hence my confidence that it can be enhanced for more widespread use. There is a great deal of activity in this direction in machine learning and computational linguistics right now. ❐

INTERVIEW BY IULIA GEORGESCU